AD_____

GRANT NUMBER: DAMD17-94-J-4015

TITLE: Digital Image Database with Gold Standard and Performance Metrics for Mammographic Image Analysis Research

PRINCIPAL INVESTIGATOR: Kevin W. Bowyer, Ph.D.

CONTRACTING ORGANIZATION: University of South Florida
Tampa, FL 33620-7900

REPORT DATE: July 1996

TYPE OF REPORT: Annual

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 1996	Annual (1 Jul 95 - 30 Jun 96)

4. TITLE AND SUBTITLE
Digital Image Database with Gold Standard and Performance Metrics for Mammographic Image Analysis Research

5. FUNDING NUMBERS
DAMD17-94-J-4015

6. AUTHOR(S)
Kevin W. Bowyer, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of South Florida
Tampa, FL 33620-7900

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, MD 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200*

The Digital Database for Screening Mammography (DDSM) is an infrastructure resource for the mammographic image analysis research community. The purpose of the DDSM is to make it possible for researchers to rigorously compare the performance of different image analysis techniques. An initial volume of DDSM has been made publicly available. A volume corresponds to an amount of data that would fit on one exabyte tape for purposes of transfer through the mail. This volume contains approximately 120 normal cases. Over 30 distinct sites around the world have already accessed (portions of) this data over the internet. In addition to the image data, a software utility for viewing the images has been made available.

14. SUBJECT TERMS
Breast Cancer

15. NUMBER OF PAGES
13

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the US Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

KWB Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

K W Bauer 29 July '96

PI - Signature Date

Contents

1 Introduction

The Digital Database for Screening Mammography (DDSM) is an infrastructure resource for the mammographic image analysis research community. The purpose of the DDSM is to make it possible for researchers to rigorously compare the performance of different image analysis techniques. The DDSM will eventually contain approximately 3,000 cases. Abnormal cases will have associated specifications of the type and location of abnormality. Software tools will be available to view the data, and to compute standardized performance metrics for detecting abnormalities.

Previously, most research on computer image analysis for mammogram screening has used a "small" (10s to perhaps 100) number of images. Also, researchers have generally not been able to evaluate their work using the same images used by other researchers. The infrastructure resource created through this project is meant to address these problems.

2 The DDSM Infrastructure Resource

This section outlines the conceptual organization of the database in terms relevant to the context of a screening program.

2.1 Database organized as cases

At the highest level, the database is organized as a set of individual cases, where a "case" is defined as a standard screening exam of two images of each breast, plus selected additional non-image information. Our plans originally called for the cases in DDSM to be divided across five broad categories of result: (1) *clearly normal*, (2) *normal after recall*, (3) *abnormal – benign*, (4) *abnormal – cancer* and (5) *false negative*. We have revised our plans to have the category *abnormal – benign* split into two categories: *abnormal – benign by mammographic criteria* and *abnormal – benign by biopsy or other*. The reason for this change is to better reflect the major types of results in the context of a screening program. The planned approximate number of cases in each category is as follows:

clearly normal	800
normal after recall	200
abnormal – benign by mammo criteria	200
abnormal – benign by biopsy or other	800
abnormal – cancer	1,000
false negative	≈ 10

2.2 Definition of "clearly normal"

This category is defined as cases presented for screening which:

- were read as normal from the standard screening exam of two views of each breast, **and**

- had a subsequent exam four years later (plus/minus six months) which was also read as normal with no more work-up than additional views, **and**
- for which there is no clinical evidence of malignancy.

Note that the *early* case, not the follow-up case, is the "clearly normal" case that goes into the database. A case which falls into this category has, by this definition, no suspicious region in any of the four images.

2.3 Definition of "normal after recall"

This category is defined as cases presented for screening which:

- were read as normal only after the reading of additional views beyond the standard screening exam, **and**
- had no need of follow-up other than the additional views, **and**
- have had at least four years of subsequent negative screening exams.

A case which falls into this category has, by this definition, at least one suspicious region in at least one image. Each image which contains a suspicious region has an associated "overlay" which records the location and type of the region.

2.4 Definition of "abnormal – benign by mammo criteria"

This category is defined as cases presented for screening in which the breast image presents at least one region that is not "normal," but is also clearly not a cancer. A case in this category has at least one suspicious region in at least one image, and each such image has an associated overlay file. Cases in this category have the four-year normal follow-up as described in the "clearly normal" definition.

2.5 Definition of "abnormal – benign by biopsy or other"

This category is defined as cases presented for screening which contained a suspicious area that was determined to be benign on the basis of biopsy, or the clear demonstration of a cyst by ultrasound or aspiration. A case in this category has at least one suspicious region in at least one image, and each such image has an associated overlay file.

2.6 Definition of "abnormal – cancer"

This category is defined as cases presented for screening which contained a suspicious area that was determined to be cancer on the basis of biopsy. A case in this category has at least one suspicious region in at least one image, and each such image has an associated overlay file.

2.7 Definition of "false negative"

This category is defined as cases presented for screening which were initially read as *clearly normal* or as *normal after recall*, but were later determined to have a cancer present. This category is further sub-divided into cancers which are (a) "clear in retrospect," and (b) "not clear in retrospect." The images for a case in this category may or may not have an associated overlay file.

3 Digitization of films

Based on current knowledge, available technology, and budget constraints at the time that the project began, the decision was made to begin accumulating studies for the database using a digitizer capable of a spatial resolution of 21 microns,[1] with an intensity resolution of 16 bits. Accordingly, a model M2100 digitizer from DBA Systems was purchased and installed at Massachusetts General Hospital.

We have experienced a number of problems in attempting to use the M2100 system in the manner required for the project. The workflow which would allow us to digitize the volume of film images needed in this project calls for loading the stack feeder with a "large number" (e.g., 20 to 28) of films, to be digitized at 21 microns and the data exported to the disk on a SUN workstation. In attempting to use the M2100 system in this manner, a number of situations were experienced which have resulted in revisions of the M2100 system by DBA. These situations range from pure "hardware" problems (e.g., rapid bulb burnout, addressed through the installation of a better ballast), through mixed hardware/software problems (e.g., unreliable film edge sensing, addressed through time-smoothing of signal), to pure software problems (e.g., too-slow communications speed using Windows-95 features, addressed through use of lower-level system features). A complete accounting of all of the situations encountered and their current state of resolution is beyond the scope of this document. However, the time spent in addressing these problems has caused the project to temporarily slip behind the original projected schedule.

We have so far digitized approximately 200+ cases worth of data. Film-screen mammograms are digitized at 21 or 42 micron resolution. (The same digitizer will not necessarily be used for all studies entered into DDSM.) Films drawn from the film library are reviewed for quality control purposes at MGH before being digitized. Each digitized image is reviewed at USF from a quality control perspective, to catch errors that can occur in the digitization process.

The "raw" image files are huge by almost any standard. Two steps are taken to reduce storage requirements. First, each image is cropped to eliminate large portions of blank background. Second, the images are stored in a compressed form using lossless JPEG. Lossless JPEG was chosen over "compress" and "gzip" on the basis of a simple comparison using three cropped, 42-micron images. The typical size of an uncompressed image is 15 MB. On average, compress reduced storage requirements by 28%, gzip by 36% and lossless JPEG by 52%. Compression speeds of the three utilities were similar. Speed of uncompression varied more substantially, and

[1]The grant proposal specified a spatial resolution at least as fine as 50 microns. We felt that, to the extent possible within budget and workload, it would be important to have at least some subset of the database digitized at the finest spatial resolution feasible.

lossless JPEG was the fastest.

4 DDSM availability to the research community

The first public release of data as part of DDSM was made in June of 1996. The announcement circulated to various electronic mailing lists appears in an appendix. Announcement was also made at the Third International Workshop on Digital Mammography [1]. This initial portion of the DDSM is available via anonymous ftp, or through mail of an exabyte tape.

The DDSM is organized as a set of cases. Each case consists of a header file, four large image files, and a small "thumbnail" image file. The header file contains patient information such as age and ACR breast density rating [3]. When abnormal cases are added, they will contain information on the type and location of lesion. The header file also contains information about the digitization process and the image files. The large image files represent the individual images for the standard left and right CC and MLO views of a screening exam. The small "thumbnail" image file represents a reduced-resolution image of the four views displayed together.

While some films are digitized at 21 microns, the resolution of the images in the on-line version of the database is 42 microns. (The 21 micron version of the data will also be made available, but at least initially only on tape.)

The directory (`pub/DDSM`) for the DDSM contains a file named `DDSM_README`, and three subdirectories named `papers`, `cases`, and `software`. The `cases` subdirectory currently contains one subdirectory, `normals`. Additional subdirectories will eventually exist at this level for the different types of studies. The `normals` subdirectory contains a subdirectory for each "volume" of normal studies, where a "volume" is meant to be a set of cases that can be distributed on one exabyte tape.

The `normal_01` volume contains a subdirectory for each of over 100 cases. Each of these subdirectories has a name of the form `institution_case_category`, where institution is a one-character code (A, B, C, ...) designating an institution, case is a four-character ascii code (0001, 0002, ...) numbering the cases for a given institution, and category is a one-character ascii code indicating to which category the case belongs. A typical actual case directory name is `A_0001_1`. For the normal category of cases, there are six files in a case directory. These files all have the same basic case name, but are distinguished by different extensions. For the example just given, the file names would be

`A_0001_1.ICS`, the header file containing information about the case,
`A_0001_1.COMB.16_PGM`, a thumbnail overview of the four images,
`A_0001_1.LEFT_CC.LJPEG`, the compressed left cranio-caudal image,
`A_0001_1.LEFT_MLO.LJPEG`, the compressed left medio-lateral image,
`A_0001_1.LEFT_MLO.LJPEG`, the compressed left medio-lateral image,
`A_0001_1.RIGHT_CC.LJPEG`, the compressed right cranio-caudal image, and
`A_0001_1.RIGHT_MLO.LJPEG`, the compressed right medio-lateral image.

The purpose of the `.PGM` image file is to facilitate "browsing" the case on a coarse-resolution, 16-bit/pixel version of the four images. This image is not suitable for use in any image processing algorithm for detection of abnormalities. Additional detail on the file formats is contained in documents available on the internet site. (This header and image file format is modeled after the international cytometry standard [2].)

The `software` directory contains a statically-linked executable version of a viewing utility named `DDSMview`. This program has been developed on SUN workstations running the SUN OS and Solaris OS, with X-windows and Motif. It should allow users to load and view cases in a reasonable fashion. Also in the `software` directory is the `C` source code for an `ics2pgm` utility. This program can be used to convert the image files of a case into (16-bit) `pgm` format. This is a format readable by, for example, the popular `xv` utility. The source code of `ics2pgm` also serves as working example of how to read the image files in DDSM. The `software` directory also contains information on the lossless JPEG utility.[2]

We are supporting access to the database via ftp over the internet and via 8 mm "exabyte" tapes sent through the mail. A list of ftp accesses to DDSM in the three weeks following the announcement of the first increment of data being made available appears in the second appendix.

We are also continuing to provide ftp access to an existing database of images provided by Nico Karssemeijer and used in his published work on algorithms for the detection of microcalcifications [4,5].

5 Conclusions

An initial volume of DDSM has been made publicly available. (A volume corresponds to an amount that would fit on one exabyte tape for purposes of sending the data through the mail.) This volume contains approximately 120 normal cases. Over 30 distinct sites around the world have already accessed (portions of) this data over the internet.

Work is proceeding to make additional "volumes" of data available in the next few months, including abnormals.

Due to problems with the digitizer, the project is currently behind on the number of cases digitized. However, several options are being pursued simultaneously to address the backlog. One option is simply to have a "night shift." A stack of films left at 5:00 to digitize overnight finishes before the next morning. A person from the night shift in the Radiology department at MGH has been identified to load a second stack of films into the digitizer. This option substantially increases our potential throughput. We are also discussing two possible options with DBA: (1) having some films digitized for the project by DBA at their Melbourne, Florida site, or (2) having a second M2100 system loaned to MGH for a period of time. Through some combination of these options, we hope to have the project back on the original schedule by the end of year three.

[2]Note that this is the true lossless JPEG algorithm, and not the more widely-known lossy JPEG algorithm set for minimum loss. The lossless JPEG standard is basically a 2-D differential predictive encoding approach.

6 References

1. Bowyer, K., Kopans, D., Kegelmeyer, W.P., Moore, R., Sallam, M, Chang, K., and Woods, K. The digital database for screening mammography, *Third International Workshop on Digital Mammography*, Chicago, June 1996.

2. P. Dean, L. Mascio, D. Ow, D. Sudar and J. Mulkin, Proposed Standard for Image Cytometry Data Files, *Cytometry* **11**, 1990, pages 561-569.

3. D.B. Kopans, C.J. D'Orsi, D.D. Adler, L.W. Bassett, R.J. Brenner, G.D. Dodd, S.A. Feig, M.A. Lopiano, R.McLelland, M. Moskowitz, E.A. Sickles. Breast Image Reporting and Data System. American College of Radiology. May 1993.

4. N. Karssemeijer. Adaptive Noise Equalization and Image Analysis in Mamography. Proceedings Information Processing In Medical Imaging (IPMI), Flagstaff, June 1993.

5. N. Karssemeijer. Adaptive Noise Equalization and Recognition of Microcalcification Clusters in Mammograms. Special issue of 'Int. Journal of Pattern Recognition and Image Analysis' on digital mammography, Vol. 7, No. 6, 1993.

A Electronic announcement of DDSM availability

ANNOUNCEMENT: VOLUME 1 OF DDSM IS AVAILABLE!!!

WHAT IS DDSM?
DDSM is the Digital Database for Screening Mammograms. Eventually the database will have 3,000 cases.

WHAT VOLUME 1 OF DDSM INCLUDES:
A volume is defined as the amount of information that fits of one 160mXL 8mm Data Cartridge. Volume 1 consists of 100+ normal 42-Micron cases where a case consists of four images. Each case contains an ACR breast density rating given by an experienced radiologist.
Also included are:
1. A README file explaining "everything" about the database.
2. Software for the viewing and conversion of the images.

This volume contains only normal cases. Future volumes will contain abnormal cases.

OBTAINING VOLUME 1:
Volume 1 can be obtained either through anonymous FTP to:
figment.csee.usf.edu in the directory pub/DDSM

or by ordering a tape directly from USF.
The tape is an EXABYTE 160mXL data cartridge. The information was written to the tape using Unix tar command.

To order a tape, please send a $30.00 check payable to University of South Florida Foundation to:
Rachel Gadsden
University of South Florida
Department of Computer Science
4202 E. Fowler Avenue
ENB 118
Tampa, FL 33620-5399

Unfortunately, we are not set up to accept purchase orders or credit cards.

Please direct any questions to: ddsm@bigpine.csee.usf.edu

For more information, please visit our web site at:
http://marathon.csee.usf.edu/announce.html

B Extract From FTP Access Log for July 1996

The following is an abstract from the log of ftp accesses made to the address made to the DDSM data for approximately a three-week period following the announcement of the availability of the first volume of normal cases. All but the first entry for any given login id have been deleted. (In many cases, one login id accounted for a large number of distinct login sessions on different days, with each session resulting in many transaction entries in the log file.) The internet addresses in the log file reveal a wide breadth of military, commercial, government and university institutions in the United States, as well as many accesses from outside the US.

1. Wed Jul 3 15:42:33 1996 1 MAMMO.PNDR.UPENN.EDU 4508 /pub/DDSM/DDSM_README a _ o a toto@mammo.pndr.upenn.edu ftp 0 *
2. Wed Jul 3 16:09:59 1996 1 kronecker.mts.jhu.edu 4508 /pub/DDSM/DDSM_README b _ o a priebe@kronecker.mts.jhu.edu ftp 0 *
3. Wed Jul 3 16:21:59 1996 1 elazteco.eecs.wsu.edu 4508 /pub/DDSM/DDSM_README b _ o a mozilla@ ftp 0 *
4. Wed Jul 3 16:28:29 1996 1 attos.bctel.net 4508 /pub/DDSM/DDSM_README b _ o a Netscape@ ftp 0 *
5. Wed Jul 3 16:29:41 1996 1 uckbv5.ece.uc.edu 4508 /pub/DDSM/DDSM_README a _ o a ychitre@ ftp 0 *
6. Wed Jul 3 16:33:07 1996 1 foehn.univ-bpclermont.fr 4508 /pub/DDSM/DDSM_README b _ o a mozilla@ ftp 0 *
7. Wed Jul 3 17:42:01 1996 1 wwwproxy2.ca.sandia.gov 4508 /pub/DDSM/DDSM_README b _ o a httpd@wwwproxy.ran.sandia.gov ftp 0 *
8. Thu Jul 4 02:58:43 1996 1 toros.ee.bilkent.edu.tr 4508 /pub/DDSM/DDSM_README b _ o a mozilla@ ftp 0 *
9. Thu Jul 4 03:20:39 1996 1 thomas-n.cevis.uni-bremen.de 4508 /pub/DDSM/DDSM_README b _ o a mozilla@ ftp 0 *
10. Thu Jul 4 04:35:27 1996 1 pc-mo035.derby.ac.uk 4508 /pub/DDSM/DDSM_README b _ o a mozilla@ ftp 0 *
11. Thu Jul 4 04:37:08 1996 1 marduk.biomed.abdn.ac.uk 4508 /pub/DDSM/DDSM_README a _ o a uns_drill@biomed.abdn.ac.uk ftp 0 *
12. Thu Jul 4 04:49:33 1996 1 sorrel.hensa.ac.uk 4508 /pub/DDSM/DDSM_README b _ o a Netscape@ ftp 0 *
13. Thu Jul 4 04:56:29 1996 1 psypc72.psy.surrey.ac.uk 4508 /pub/DDSM/DDSM_README a _ o a p.sowden@surrey.ac.uk ftp 0 *
14. Thu Jul 4 06:13:34 1996 1 genmin85.ee.wits.ac.za 4508 /pub/DDSM/DDSM_README b _ o a yelland@odie.ee.wits.ac.za ftp 0 *
15. Thu Jul 4 06:28:21 1996 1 erikthur.anst.uu.se 4508 /pub/DDSM/DDSM_README b _ o a mozilla@ ftp 0 *
16. Thu Jul 4 13:31:49 1996 1 orpheus.ecn.purdue.edu 4508 /pub/DDSM/DDSM_README b _ o a chiren@ ftp 0 *

17. Fri Jul 5 01:23:03 1996 1 i3a.ua.es 4508 /pub/DDSM/DDSM_README a _ o a sco@dtic.ua.es ftp 0 *
18. Fri Jul 5 06:34:06 1996 1 tilki4.nswc.navy.mil 4508 /pub/DDSM/DDSM_README b _ o a dmarche@tilki4.nswc.navy.mil ftp 0 *
19. Sat Jul 6 13:32:54 1996 1 pine.lle.rochester.edu 4508 /pub/DDSM/DDSM_README a _ o a ajr@ ftp 0 *
20. Wed Jul 10 06:42:43 1996 1 mgi_jkn_pc.medphys.ucl.ac.uk 4508 /pub/DDSM/DDSM_README b _ o a mozilla@ ftp 0 *
21. Wed Jul 10 09:19:20 1996 1 host139.cat.pinellas.k12.fl.us 4508 /pub/DDSM/DDSM_README b _ o a mozilla@ ftp 0 *
22. Wed Jul 10 11:11:15 1996 1 rsd1000.gtri.gatech.edu 4508 /pub/DDSM/DDSM_README a _ o a chris.barnes@gtri.gatech.edu ftp 0 *
23. Thu Jul 11 12:29:02 1996 1 santorini.ee.ic.ac.uk 365 /pub/DDSM/software/src/ics2pgm/ics2pgm _README b _ o a mozilla@ ftp 0 *
24. Thu Jul 11 13:37:03 1996 1 enricbas.socio.ua.es 100 /pub/DDSM/software/doc/DDSMView.doc b _ o a bas@ua.es ftp 0 *
25. Tue Jul 16 11:30:56 1996 1 etoile.igd.u-bordeaux.fr 4508 /pub/DDSM/DDSM_README b _ o a mozilla@ ftp 0 *
26. Wed Jul 17 19:27:50 1996 1 conciliator.acsu.buffalo.edu 4508 /pub/DDSM/DDSM_README b _ o a mozilla@ ftp 0 *
27. Thu Jul 18 16:19:29 1996 1 wednesday.polaroid.com 4508 /pub/DDSM/DDSM_README a _ o a scarffl@polaroid.com ftp 0 *
28. Fri Jul 19 07:58:42 1996 2 esg_735.dibe.unige.it 4508 /pub/DDSM/DDSM_README b _ o a zunino@dibe.unige.it ftp 0 *
29. Fri Jul 19 14:50:10 1996 1 interlock.arinc.com 4508 /pub/DDSM/DDSM_README b _ o a jweingar@arinc.com ftp 0 *
30. Sun Jul 21 20:39:48 1996 1 internet.archive.org 4508 /pub/DDSM/DDSM_README b _ o a Robert@archive.org ftp 0 *
31. Sun Jul 21 22:37:42 1996 1 pm1-ppp64.well.com 4508 /pub/DDSM/DDSM_README b _ o a mozilla@ ftp 0 *
32. Mon Jul 22 20:51:05 1996 1 fornax.cis.umassd.edu 100 /pub/DDSM/software/doc/DDSMView.doc a _ o a gwolee@cis.umassd.edu ftp 0 *
33. Tue Jul 23 22:03:51 1996 1 inet-tis.toshiba.co.jp 4508 /pub/DDSM/DDSM_README b _ o a Netscape@ ftp 0 *
34. Wed Jul 24 14:13:06 1996 1 dagwood.bme.ri.ccf.org 4508 /pub/DDSM/DDSM_README b _ o a makela@bme.ri.ccf.org ftp 0 *